Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack

Year: 2023 | Citations: 263 | Authors: Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam S. Tsai, Jialiang Wang

---

Abstract

Training text-to-image models with web scale image-text pairs enables the generation of a wide range of visual concepts from text. However, these pre-trained models often face challenges when it comes to generating highly aesthetic images. This creates the need for aesthetic alignment post pre-training. In this paper, we propose quality-tuning to effectively guide a pre-trained model to exclusively generate highly visually appealing images, while maintaining generality across visual concepts. Our key insight is that supervised fine-tuning with a set of surprisingly small but extremely visually appealing images can significantly improve the generation quality. We pre-train a latent diffusion model on 1.1 billion image-text pairs and fine-tune it with only a few thousand carefully selected high-quality images. The resulting model, Emu, achieves a win rate of 82.9% compared with its pre-trained only counterpart. Compared to the state-of-the-art SDXLv1.0, Emu is preferred 68.4% and 71.3% of the time on visual appeal on the standard PartiPrompts and our Open User Input benchmark based on the real-world usage of text-to-image models. In addition, we show that quality-tuning is a generic approach that is also effective for other architectures, including pixel diffusion and masked generative transformer models.